基于多策略的临床术语标准化

林楠铠[1] 林晓钿[2] 吴凯莹[3] 陈枫[2] 蒋盛益[2,4]

（1.广东工业大学 计算机学院，广东省 广州市 510000；2.广东外语外贸大学 信息科学与技术学院，广东省 广州市 510000；3. 广东外语外贸大学 数学与统计学院，广东省 广州市 510000； 4. 广东外语外贸大学 广州市非通用语种智能处理重点实验室，广东省 广州市 510000）

摘要：临床术语标准化对于处理电子病历中临床术语不规范问题具有重要的研究意义。目前主流的解决方法是采用“召回-排序”的策略。该文基于中国健康信息处理大会（CHIP2021）[1]评测 3 中提供的数据集，提出了一个基于多策略的临床术语标准化方法，在召回阶段，采用全匹配策略、相似原词的标准词推荐以及基于 TF-IDF 与改进的 Jaccard 系数的相似度计算去召回候选的标准词集合。同时，该文构建了基于 BERT 模型的标准词数量预测模型，利用对抗训练、Focal Loss 与标签平滑策略有效地提高了模型的预测性能和泛化性能。在排序阶段，该文利用基于对抗训练与诊断信息融合的 BERT 蕴含分数排序模型对候选词集合排序，再根据数量预测模型输出的结果生成最终预测的标准词。在最终的评测中，该文方法准确率达到 0.6356，在参赛队伍中位列第二名。

关键词：术语标准化；多策略融合；相似度计算

中图分类号：TP391 **文献标识码：**A

Clinical Term Normalization Based on Multiple Strategies

Nankai Lin[1], Xiaotian Lin[2], Kaiying Wu[3], Feng Chen[2] and Shengyi Jiang[1,4]

(1. School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, Guangdong, 510000, China; 2. School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, Guangdong, 510000, China; 3. School of Mathematics and Statistics, Guangdong University of Foreign Studies, Guangzhou, Guangdong, 510000, China; 4. Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies, Guangzhou, Guangdong, 510000, China)

Abstract: The clinical term normalization has important research significance for dealing with the problem of non-standardization of clinical terminology in electronic medical records. The current mainstream solution is to adopt a "recall-sort" strategy. Based on the dataset provided in Evaluation 3 of the China Conference of Health Information Processing, we propose a multi-strategy-based normalization method for clinical terms. In the recall phase, the full-matching strategy, standard words recommendation of similar original words, and similarity calculation based on the TF-IDF and the improved Jaccard coefficient are used to recall the candidate standard word set. At the same time, we construct a standard quantity prediction model based on the BERT model, and use adversarial training, focal loss and label smoothing strategies to effectively improve the prediction performance and generalization performance of the models. In the ranking stage, In the ranking stage, we use the BERT implicit score ranking model based on adversarial training and fusion of diagnostic information to rank the candidate word set, and then generate the final predicted standard words based on the output of the quantity prediction model. In the final evaluation test set, the method accuracy rate of our method reached 0.6356, ranking second place among the participating teams.

Key words: Clinical Term Normalization; Multiple Strategies; Similarity Calculation

[1] 中国中文信息学会医疗健康与生物信息处理专业委员会第七届中国健康信息处理大会（http://cips-chip.org.cn，2021.12.04-05）

0 引言

由于医护人员个人表述习惯的不同，在临床实践中，同一种诊断、手术、药品、检查、化验、症状等表述会有成百上千种不同的写法。在医疗信息处理领域中，临床术语标准化任务即是为临床上各种不同说法找到对应的标准表述。

临床医学术语标准是面向计算机应用、以概念为中心的术语一体化系统，以疾病诊断为核心，涵盖身体结构病因、病理、临床表现、临床诊断技术与方法、操作技术、医学仪器与设备、护理、社会背景、物理等范畴。智能诊疗、智能影像识别等在实施过程中面临临床医学术语不规范、临床医学知识匮乏、不全面、不成体系以及词法、句法、语义、语用存在不确定性等问题，需要统一、规范的术语标准体系作为支撑，实现各系统底层数据的标准化、规范化。而临床医学术语标准的应用，可用来精确表达医学概念，编码、提取和分析医学数据，支持医学数据的一致性索引、存储、调用和跨系统集成，实现医疗数据的语义互操作，在医疗人工智能领域发挥重要作用。

语义关系又称语义结构、语义结构关系，是词语概念意义间的关系抽象概括的结果。医疗人工智能要实现一定程度上的智能自主性、独立性，必须有能力对数据的语义关系进行关联和处理。中文临床医学术语标准利用相似度、疑似性、深度学习等算法处理自然语言，深度挖掘潜在语义关系，实现疾病/诊断与发病部位、临床表现、临床观察、检查、治疗、病理、化学品、药物品、形态学等具有临床意义的不同医疗元素和相关元素之间的语义关联，为医疗人工智能的运作机制提供基于事实的计算语境，对破解医疗人工智能实践难题起到支撑性作用。

规范化的临床医学术语标准可以消除临床概念的不确定性，以支持医疗数据的精确记录与分析；实现不同系统间医疗数据的分享与利用；促进人工智能与医疗健康领域的深度融合。

2013 年，ShARe/CLEF eHealth[1]最早发布了英语临床术语标准化数据。此外，在 SemEval-2014 Task 7[2]和 SemEval-2015 Task 14[3]发布了英语临床术语标准化的评测任务。第五届中国健康信息处理会议（CHIP 2019）[4]发布了中文临床术语标准化任务与数据集，推动了该任务在中文自然语言处理上的发展，然而，目前中文临床术语标准化还处在起步阶段，相关研究也较少。

2021 年，第七届中国健康信息处理会议（CHIP 2021）开放了第三届中文临床术语标准化评测任务，本团队参加该比赛并获得第二名，后续本文将从中文临床术语标准化任务的研究现状，CHIP 2021 的数据描述与本团队构建的临床术语标准化模型描述三个方面展开，并对实验结果进行分析与总结展望。

1 相关研究

早期面向临床术语标准化任务采用基于规则的方法。Ghiasvand 等[5]采用基于编辑距离特征的方法生成候选集，通过训练集学习到 554 种编辑距离模式，在 SemEval-2014 任务 7 上取得了最佳效果。Kang 等[6]提出了 5 种规则来提升疾病术语的归一化性能。

目前，针对临床术语标准化任务大多数采用“召回-排序”的策略。Leaman 等[7]首次提出了一种成对（pairwise）学习排序技术，该技术采用矢量空间模型来计算非标准化医学实体和标准化医学实体两者的文本相似度。Luo 等[8]提出一个多任务框架，可以对疾病和手术操作类实体进行规范化，多任务共享结构使模型能够利用疾病和手术操作之间的医学相关性，更好地执行消歧任务。Ji 等[9]通过微调的预训练 BERT 模型来实现实体规范化。在中文临床术语标准化任务上，崇伟峰等[10]基于文本蕴含的思想，构建了临床术语标准化系统，由数据预处理、BERT 蕴含打分、BERT 数量预测以及基于逻辑回归的重排序四个模块组成，在第五届中国健康信息处理大会评测 1 的测试集达到了 94.825%的性能，评测排名第一。陈沫沙等[11]设计了两种检索方式，通过检索“编码-标准词”与“标注历史”得到候选标准词，再基于文本蕴含的思想对候选标准词进行重排序，其提出的方法在测试集上单模型达到了 89.1%、融合模型达到 92.8%的性能。孙日君等[12]提出了一种基于 BERT 的临床术语标准化方法。该方法使用 Jaccard 相似度算法从标准术语集中挑选出候选词，基于 BERT 模型对原始词和候选词进行匹配得到标准化的结果，该方法在测试集上准确率为 90.04%。与孙日君等人相似，杨飞洪等[13]通过融合文本相似度排序+BERT 模型匹配开展建模，该

方法在第五届中国健康信息处理大会评测 1 的测试集准确率为 88.51%。

除了采用“召回-排序”的策略，还有学者尝试将生成式方法的思想应用于该任务。闫璟辉将临床术语标准化任务类比为翻译任务，引入深度生成式模型对描述文本的核心语义进行生成并得到标准词候选集，再利用基于 BERT 的语义相似度算法对候选集进行重排序得到最终标准词。

2 基于多策略的临床术语标准化方法

如图 1 所示，本文提出了一个基于多策略的临床术语标准化方法，包括候选标准词召回模块、标准词数量预测模块和候选标准词排序模块。在召回阶段，采用全匹配策略、相似原词的标准词推荐以及基于 TF-IDF 与改进的 Jaccard 系数的相似度计算去召回候选的标准词集合。同时，本文构建了基于 BERT 模型的标准词数量预测模型，利用对抗训练、Focal Loss 与标签平滑策略有效地提高了模型的预测性能和泛化性能。在排序阶段，本文利用基于对抗训练与诊断信息融合的 BERT 蕴含分数排序模型对候选词集合排序，再根据数量预测模型输出的结果生成最终预测的标准词。

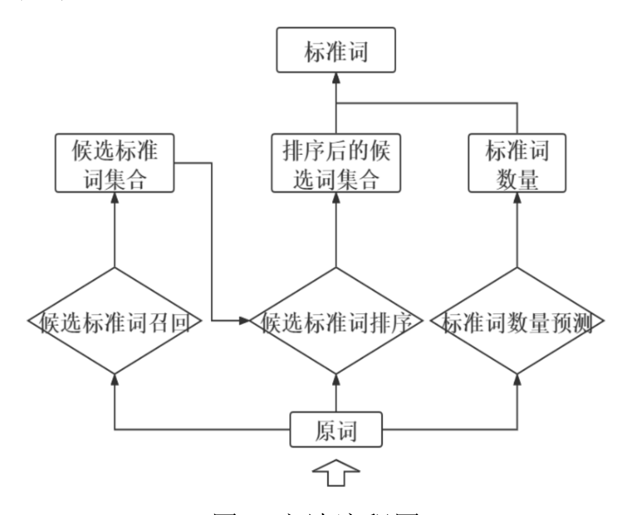


图 1 方法流程图

2.1 候选标准词召回

在召回阶段，采用全匹配策略、相似原词的标准词推荐以及基于 TF-IDF 与改进的 Jaccard 系数的原词-标准词相似度计算去召回候选的标准词集合。假设 ICD-10 标准词列表为$S=\{S_1, S_2, \dots, S_n\}$，训练集样本为$X=\{X_1, X_2, \dots, X_n\}$，样本$X_i$（$X_i \in X$）的原词为$O_i$，候选标准词列表为$H_i$。

本文首先采用全匹配策略筛选候选标准词，对于样本X_i，遍历标准词列表S，判断每一个S_j（$S_j \in S$）字符串是否完全出现在字符串O_i中，若完全出现在字符串O_i中，则添加到候选标准词列表H_i。

在相似原词的标准词推荐模块中，本文基于X中的所有原词训练一个 TF-IDF 模型M_O，通过该模型，计算了每一个样本X_i的原词O_i与X中的其它样本的原词之间的余弦相似度，对于样本X_i，选择相似度最高的k个相似原词，将相似原词所对应的标准词添加到候选标准词列表H_i中。实验过程中，k取 40。

在基于 TF-IDF 与改进的 Jaccard 系数的原词-标准词相似度计算模块中，利用 ICD-10 标准词列表S训练一个 TF-IDF 模型M_s，通过该模型计算样本X_i的原词O_i与标准词S_j之间的余弦相似度Sim_T。此外，还计算了原词O_i与标准词S_j之间的字级别 Jaccard 系数。原始的 Jaccard 系数表示为：

$$Jaccard(O_i, S_j) = \frac{|O_i \cap S_j|}{|O_i \cup S_j|}$$

由于原词中可能包含多个标准词，原词中包含的信息量远远大于每个标准词中包含的信息量，从而导致 Jaccard 系数普遍偏低。因此，本文在计算 Jaccard 系数时，分母只考虑标准词中出现的字的数量，即

$$Jaccard(O_i, S_j) = \frac{|O_i \cap S_j|}{|S_j|}$$

通过改进后的 Jaccard 系数计算原词O_i与标准词S_j之间的相似度为Sim_j，则原词O_i与标准词S_j之间的融合相似度为：

$$S = Sim_T + Sim_j$$

通过两种相似度计算方法融合得到O_i与S中的每个标准词之间的相似度，根据相似度排序并选择前r个标准词添加到候选标准词列表H_i中。实验过程中，r取 100。

通过三种步骤召回的候选标准词可能出现重复现象，经过去重后得到的候选标准词列表即为最终的候选标准词列表。

2.2 标准词数量预测

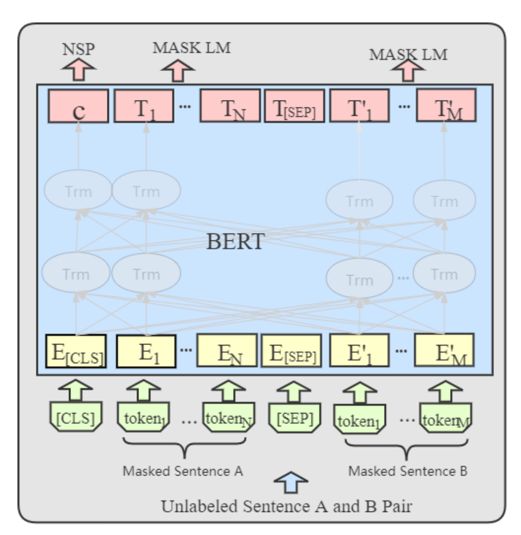


图 2 BERT 模型

本文基于 BERT 模型（如图 2 所示）构建了标准词数量预测模型，同时融合了不同的策略提升模型的预测性能和泛化性能，评测阶段，融合三种策略下的模型：基于对抗训练的 BERT 模型、基于对抗训练与 Focal Loss 的 BERT 模型以及基于对抗训练与标签平滑的 BERT 模型。标准词数量预测模型的标签分为三类：只含有一个标准词、包含两个标准词以及包含两个以上的标准词。

2.2.1 对抗训练

本文利用对抗训练增加模型的多样性和泛化性，采用 Fast Gradient Method（FGM）对模型的 Embedding 层注入噪声进行扰动，注入的扰动定义为：

$$r_{adv} = \alpha \cdot g/\|g\|_2$$

$$g = \nabla_x L(x,\ y;\theta)$$

其中g为原始的模型梯度，α为句子词汇的权重矩阵，本文实验过程中设置为 1。扰动后的梯度为：

$$g' = r_{adv} + g$$

对抗训练过程中，本文采用扰动后的梯度进行反向传播和参数更新后，再将 Embedding 层的噪声移除，恢复原来的梯度，进行下一轮的迭代训练。

2.2.2 Focal Loss

对于每一个类别，由于负样本多于正样本，因此分类结果可能会存在偏差。本文除了采用交叉熵作为损失函数构建模型之外，还采用了 Focal Loss 来帮助缓解类不平衡的问题。这个损失函数旨在减少训练中大量的简单负样本的权值。Focal Loss 的公式如下：

$$L = \sum_{i=1}^{C} (1-p_i)^{\gamma} \log(p_i)$$

其中权重系数 γ 是超参数。Lin 等[15]验证了 γ 的最优值是 2，本文实验也将 γ 的值设为 2。

2.2.3 标签平滑

对于交叉熵损失函数，在模型训练阶段需要使用预测概率来拟合真实概率。但是，模型拟合 one-hot 编码的标签会导致其预测结果对于真实标签的过拟合，因此无法保证模型的泛化能力。本文采用标签平滑技术来减轻模型的过度拟合的问题。假设y是一个 one-hot 编码的标签，经过标签平滑后的真实标签可以表示为：

$$y_i' = (1-\varepsilon) * y_i + \frac{\varepsilon}{|K|}$$

标签平滑后的损失为：

$$L = \sum_{i=1}^{C} y_i' * \log(p_i)$$

其中ε是平滑因子，K是类别的数量。在标准词数量预测任务中，K是 3， ε设置为 0.05。

2.3 候选标准词排序

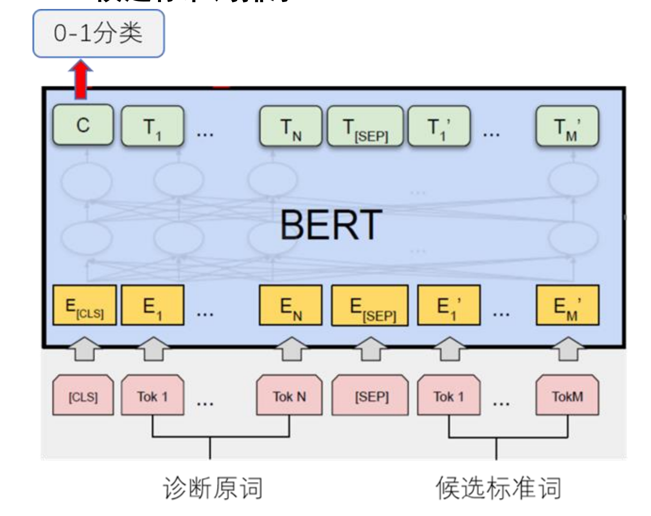


图 3 文本蕴含模型

本文基于 BERT 模型构建了文本蕴含模型（如图 3 所示），主要是对给定的诊断原词计算其与每一个标准名的蕴含分数。本文将诊断原词与候选标准词拼接后输入到 BERT 模型进行 0-1 分类，0 代表候选标准词不是该原词正确的标注词，1 则代表标准词为该原词正确的标准词，将分类为 1 的概率作为蕴含分数。

本文融合四种策略下的模型：BERT 模型、基于对抗训练的 BERT 模型、融合手术原词数据的 BERT 模型、基于对抗训练与融合手术原词数据的 BERT 模型。

在数据构建部分，针对手术原词数据，本文只将手术原词数据作为正样本，评测任务提供的原词数据样本为 2500 条，处理后的数据为 2605 条，由于该部分数据样本太少，因此本文将该部分数据进行了四倍数据量的扩充，最后共生成的手术原词训练样本为 10420 条。本文在构建诊断原词数据时，采用 2.1 中的相似原词的标准词推荐模块，提取最相似的五个相似原词的标准词以及基于 TF-IDF 模型M_S的最相似的十个标准词，将这一部分标准词中不在答案中的词作为负样本，答案中所有的标准词作为正样本。该策略下构建的正样本数量为 12984 条，负样本数量为 89688 条，因此本文将正样本进行了七倍数据扩充，最终生成的正样本数量为 90888 条。

2.4 多阶段结果融合

在 2.1 得到召回的候选标准词之后，采用 2.3 的蕴含模型对候选标准词进行排序，同时采用 2.2 的标准词数量预测模型预测测试样本对应的标准词数量，若标准词数量预测模型识别样本为只含有一个标准词或包含两个标准词，则分别推荐评分最高的一/两个标准词。如果模型预测的标准词数量大于两个，则采用预测概率大于 0.5 的标准词，若预测概率大于 0.5 的标准词数量少于三个，则选择评分最高的三个标准词。

3 数据描述

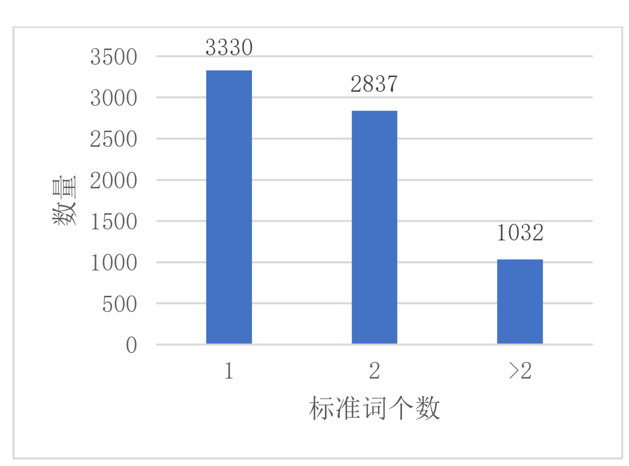


图 4 标准词个数数据分布

先将训练数据中的标准词进行预处理，去掉标记为“O”的标准词，同时去掉文本中"、"等符号。将训练集中的所有标准词与《国际疾病分类 ICD-10 北京临床版 v601》中的标准词合并去重后得到新的标准词列表，新的标准词列表共包含 37869 个标准词。预处理后的标准词数量按只含有一个标准词、包含两个标准词以及包含两个以上的标准词划分，数据分布如图 4 所示。同时，我们统计了诊断原词的一系列长度特征，结果如表 1 所示。

表 1 原词长度统计特征

属性	数值
平均值	11.314766
方差	6.281472
最小值	2
最大值	103

4 实验结果与分析

4.1 实验设置

本文的实验采用 RTX TITAN GPU 进行训练与预测，代码基于 1.7.0 的 Pytorch 框架与 4.4.0 的 Transformers 框架实现，在标准词数量预测模块与候选标准词排序模块，分别将构建的数据进行五折划分，候选标准词排序模块中的手术原词数据用于每一折的模型训练。预测阶段，每一个模型的输出结果为交叉验证的五个模型预测的概率平均值，不同模型的结果融合为各个模型的概率平均值。在候选标准词召回阶段，本文采用 strict 的准确率进行评估，当样本答案中的所有标准词都完全召回才视为正确。在标准词数量预测模块与候选标准词排序模块，各个分类模型采用准确率进行评估。在标准词数量预测模块与候选标准词排序模块中，本文采用 MedBERT-wwm[2]作为预训练的基模型，微调阶段的参数值如表 2 所示。

表 2 微调阶段参数值

模块	参数	参数值
标准词数量预测模块	Batch Size	32
	Sequence Length	32
	Epoch	15
	Learning Rate	5e-5
候选标准词排序模块	Batch Size	256
	Sequence Length	64
	Epoch	10
	Learning Rate	5e-5

4.2 候选词召回实验结果

本文首先探究了 TF-IDF 与改进的 Jaccard 系数的融合策略的有效性，分别尝试了以下几种组合：基于 TF-IDF 的相似度计算、基于 Jaccard 系

[2] https://github.com/trueto/medbert

数的相似度计算、基于改进的 Jaccard 系数的相似度计算、融合 TF-IDF 与 Jaccard 系数的相似度计算与融合 TF-IDF 与改进的 Jaccard 系数的相似度计算，实验结果如表 3 所示，单独使用时，改进的 Jaccard 系数性能略低于原始的 Jaccard 系数，准确率降低了 0.79%，融合使用时，基于改进的 Jaccard 系数与 TF-IDF 的相似度计算方法性能更好，该策略比单独使用 TF-IDF 与单独使用改进的 Jaccard 系数分别提升了 0.46%和 2%，可见融合策略可以有效地提升召回阶段的模型性能。

表 3 相似度计算方法对比实验结果

方法	准确率
TF-IDF	93.25%
Jaccard	92.50%
改进的 Jaccard	91.71%
TF-IDF + Jaccard	93.57%
TF-IDF + 改进的 Jaccard	**93.71%**

为了进一步探究全匹配策略、相似原词的标准词推荐以及相似度计算三种方法的有效性，本文进行了消融实验，结果如表 4 所示。

表 4 消融实验结果

方法	准确率
本文采用的候选词召回方法	93. 71%
-全匹配策略	92. 94%
-相似原词的标准词推荐	74. 20%
-相似度计算	72. 20%

4.3 标准词数量预测实验结果

本文探究了不同策略对标准词数量预测任务的提升，如表 5 所示，可以看到对抗训练为模型带来了有效的提升，准确率提高了 0.63%。此外，针对损失函数的处理（Focal Loss 与标签平滑），也为模型带来了稳定的性能提升。

表 5 标准词数量预测模型对比实验结果

方法	准确率
BERT	77.14%
BERT + 对抗训练	77.77%
BERT + 对抗训练 + Focal Loss	77.91%
BERT + 对抗训练 + 标签平滑	78.51%

4.4 候选标准词排序实验结果

本文采用的四种文本蕴含模型的实验结果如表 6 所示。基于对抗训练与融合手术原词数据的 BERT 模型性能最佳，模型准确率达到了 94.26%，单独融合手术原词数据时，模型性能有略微下降，在最终评测阶段，为了提高模型的泛化性，本文仍融合了该模型的结果。

表 6 候选标准词排序模型对比实验结果

方法	准确率
BERT	94.08%
融合手术原词数据的 BERT 模型	94.05%
基于对抗训练的 BERT 模型	94.22%
基于对抗训练与融合手术原词数据的 BERT 模型	94.26%

4.5 错误分析

表 7 候选词召回模块错误样例分析

错误样例 1	
原词	大汗腺炎
标准词	未特指的顶(浆分)泌汗腺疾患
未能召回的标准词	未特指的顶(浆分)泌汗腺疾患
错误样例 2	
原词	左足底黑色素瘤扩大切除术后，左下肢肌肉间转移，肺、肝转移？，胃溃疡
标准词	足部恶性肿瘤、恶性黑色素瘤、下肢继发恶性肿瘤、肺继发恶性肿瘤、肝继发恶性肿瘤、胃溃疡、转移性恶性黑色素瘤
未能召回的标准词	下肢继发恶性肿瘤
错误样例 3	
原词	前列腺癌骨髓侵犯
标准词	前列腺恶性肿瘤、骨髓继发恶性肿瘤、癌
未能召回的标准词	骨髓继发恶性肿瘤

本文进一步对候选词召回模块进行了错误分析，表 7 列举了 3 个未能完全召回所有候选词的样例，可见候选词召回模块在以下两种情况下较难正确地召回：(1)标准词与原词的相似度过低；（2）原词对应的标准词数量过多，未能完全召回全部对应的标准词。

在候选标准词排序模块中，模型容易将正确标准词的上位标准词识别为标准词，比如“左足底裂伤伴感染”的标准词之一为“足软组织感染”，

模型将原词与“足软组织感染”的上位标准词“感染”识别为蕴含关系；“双子宫双宫颈伴双阴道”的标准词为“双子宫双子宫颈双阴道”，模型将原词与“双子宫”识别为蕴含关系。上位词误判的现象明显影响了文本蕴含模型的性能。

在标准词数量预测模块中，模型难以只利用原词文本准确判断标准词数量，存在两种较极端的情况是模型基本无法准确判断的：原词文本长但只对应一个或两个标准词、原词文本长度短但包含多个标准词，如表 8 所示。

表 8 标准词数量预测模块错误样例分析

错误样例 1	
原词	左顶枕骨瘤
标准词	骨瘤、枕骨良性肿瘤、顶骨良性肿瘤
错误样例 2	
原词	左乳肿块癌
标准词	乳房肿物、乳腺恶性肿瘤、癌
错误样例 3	
原词	头面部跌伤
标准词	头部损伤、摔伤、面部损伤
错误样例 4	
原词	皮炎腰腿痛
标准词	皮炎、下肢疼痛、腰痛
错误样例 5	
原词	回肠系膜高侵袭性 B 细胞源性非霍奇金淋巴瘤 II 期（Burrkitt 淋巴瘤）化疗后
标准词	非霍奇金淋巴瘤(B 细胞型)
错误样例 6	
原词	经典型霍奇金淋巴瘤 IV 期 B 结节硬化型、侵及右颈部、双锁骨区、前纵隔、右胸肌间、右腋窝、双侧膈区、左内乳区淋巴结、侵及骨（多发）
标准词	结节硬化型经典霍奇金淋巴瘤

4. 6 最终测试集结果

在最终的评测阶段，本文在标准词数量预测模块融合了三个策略的模型，在候选标准词排序模块融合了四个策略的模型。本文的方法准确率达到 0.6356，在参赛队伍中位列第二名。

5 总结

在 CHIP 2021 评测任务 3 上，本文提出了基于多种策略的临床术语标准化方法，该方法旨在增强模型在各个阶段的泛化能力，从而提高模型的性能。在召回阶段，采用全匹配策略、相似原词的标准词推荐以及基于 TF-IDF 与改进的 Jaccard 系数的相似度计算去召回候选的标准词集合。同时，本文构建了基于 BERT 模型的标准词数量预测模型，利用对抗训练、Focal Loss 与标签平滑策略有效地提高了模型的预测性能和泛化性能。在排序阶段，本文利用基于对抗训练与诊断信息融合的 BERT 蕴含分数排序模型对候选词集合排序，再根据数量预测模型输出的结果生成最终预测的标准词。最终的评测测试集中，本文的方法准确率达到 0.6356，在参赛队伍中位列第二名。

本文提出方法仍存在一定的局限性，在预训练模型上只采用了 BERT 模型，在标准候选词排序模块中，只是单纯利用手术原词数据进行微调，并未很好利用手术原词数据的信息，未来我们将进一步改进本文的模型，尝试不同的预训练模型在该任务上的效果，同时进一步深入研究如何更好地融合与提取手术原词数据的信息辅助该任务。

参考文献

[1] Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013[C]//International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, 2013: 212-231.
[2] Sameer Pradhan, Noemie Elhadad, Wendy Chapman, et al. Semeval-2014 task 7: Analysis of clinical text[C]//Proceedings of the 8th International Workshop on Semantic Evaluation, 2014: 54-62.
[3] Noemie Elhadad, Sameer Pradhan, Sharon Gorman, et al. SemEval-2015 task 14: Analysis of clinical text[C]//proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015: 303-310.
[4] 黄源航,焦晓康,汤步洲,陈清财,闫峻.CHIP2019 评测任务 1 概述：临床术语标准化任务[J].中文信息学报,2021,35(03):94-99.
[5] Ghiasvand O, Kate R J. R.: UWM: Disorder mention extraction from clinical text using CRFs and normalization using learned edit distance patterns[C]//Proceedings of the 8th International Workshop on Semantic Evaluation, 2014: 828-832.
[6] Kang N, Singh B, Afzal Z, et al. Using rule-based natural language processing to improve disease normalization in biomedical text[J]. Journal of the American Medical Informatics Association, 2013, 20(5): 876-881.
[7] Leaman R, Islamaj Doğan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank[J]. Bioinformatics, 2013, 29(22): 2909-2917.
[8] Luo Y, Song G, Li P, et al. Multi-task medical concept normalization using multi-view convolutional neural network[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[9] Zongcheng Ji, Qiang Wei, Hua Xu. BERT-based ranking

for biomedical entity normalization[J]. AMIA Summits on Translational Science Proceedings, 2020.

[10] 崇伟峰,李慧,李雪,任禾,于东,王晔晗.基于 BERT 蕴含推理的术语标准化系统[J].中文信息学报,2021,35(05):86-90.

[11] 陈漠沙,仇伟,谭传奇.基于 BERT 的手术名称标准化重排序算法[J].中文信息学报,2021,35(03):88-93.

[12] 孙曰君,刘智强,杨志豪,林鸿飞.基于 BERT 的临床术语标准化[J].中文信息学报,2021,35(04):75-82.

[13] 杨飞洪,孙海霞,李姣.一种文本相似度与 BERT 模型融合的手术操作术语归一化方法[J].中文信息学报,2021,35(04):44-50.

[14] 闫璟辉,向露,周玉,孙建,陈思,薛晨.深度生成式模型在临床术语标准化中的应用[J].中文信息学报,2021,35(05):77-85.

[15] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.